基于改进的深度卷积神经网络的人体动作识别方法 *

陈胜娣，魏　维，何冰倩，陈思宇，刘基缘

(成都信息工程大学 计算机学院, 成都 610225)

摘　要：针对现有的动作识别算法的特征提取复杂、识别率低等问题，提出了基于批归一化变换(batch normalization)与GoogLeNet 网络模型相结合的网络结构，将图像分类领域的批归一化思想应用到动作识别领域中进行训练算法改进，实现了对视频动作训练样本的网络输入进行微批量(mini-batch)归一化处理。该方法以 RGB 图像作为空间网络的输入，光流场作为时间网络输入，然后融合时空网络得到最终动作识别结果。在 UCF101 和 HMDB51 数据集上进行实验，分别取得了 93.50%和 68.32%的准确率。实验结果表明，改进的网络架构在视频人体动作识别问题上具有较高的识别准确率。

关键词：动作识别；批归一化；深度学习；卷积神经网络

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2017.10.1017

Action recognition base on improved deep convolutional neural network

Chen Shengdi, Wei Wei, He Bingqian, Chen Siyu, Liu Jiyuan

(*College of Computer Science & Technology*, *Chengdu University of Information Technology*, *Chengdu* 610225, *China*)

Abstract: Aiming at the problem of complex feature extraction and low accuracy in human action recognition, this paper proposed a network structure combining batch normalization algorithm with GoogLeNet network model. Applying Batch Normalization idea in the field of image classification to action recognition field, it improved the algorithm by normalizing the network input training sample by mini-batch. For convolutional network, RGB image was the spatial input, and stacked optical flows was the temporal input. Then, it fused the spatio-temporal networks to get the final action recognition result. It trained and evaluated the architecture on the standard video actions benchmarks of UCF101 and HMDB51, which achieved the accuracy of 93.50% and 68.32%. The results show that the improved convolutional neural network has a significant improvement in improving the recognition rate and has obvious advantages in action recognition.

Key Words: action recognition; batch normalization; deep learning; convolutional neural network

0　引言

人体动作识别作为计算机视觉领域的一个重要研究课题，在视频监控、视频内容检索、辅助医疗、虚拟现实、智能人机交互等领域中有着广泛的应用价值和研究意义[1~4]。相比于静态图像，视频不仅具有外观信息还具有运动信息，因此动作识别的性能受到更多因素的影响，如运动场景的不同光照、视角、背景以及动作姿态的差异等。当前国内外常用的动作识别方法主要可以分为两大类：a)传统的动作识别方法；b)基于卷积神经网络的动作识别方法。

传统的动作识别方法主要是对 RGB 图像序列进行分析。申晓霞等人[5]提出结合深度信息和 RGB 图像来识别人的行为动作。张杰等人[6]提出利用时空梯度直方图和光流直方图描述子结合 K-均值的人体动作识别方法。Shotton 等人[7]利用 Harris 检测器和 Gabor 检测器来检测时空兴趣点，构建 3 维梯度直方图(HOG3D)表示特征，提出了彩色时空兴趣点的人体动作识别方法。Ofli 等人[8]提出了最大信息关节序列(SMIJ)来表示特征的识别方法。Chen 等人[9]利用正面、侧面和俯视三个投影视图中得到的深度运动图(DMMs)来捕获运动信息，再用 LBP 局部二元模式进行特征表示。赵晓健等人[10]提出结合稠密光流轨迹和稀疏编码框架的特征提取方法(DOF-SC)进行动作识别。李亚玮等人[11]提出基于单层正则化的光流约束自编码器的特征学习算法来进行动作识别。

基于卷积神经网络的动作识别方法，主要在于构建一个更有效的网络识别架构。Simonyan 等人[12]提出一种双流网络结构，证明了使用帧间光流特征训练的卷积神经网络在数据集有限的

基金项目：四川省教育厅重点科研项目（2017Z026）

作者简介：陈胜娣（1992-），女，广东阳江人，硕士研究生，主要研究方向为图形图像处理（chensheng_di@163.com）；魏维（1976-），男，教授，博士，主要研究方向为图形图像处理；何冰倩（1994-），女，硕士研究生，主要研究方向为图形图像处理；陈思宇（1998-），男，本科生，主要研究方向为图形图像处理；刘基缘（1998-），男，本科生，主要研究方向为图形图像处理.

条件下，网络依旧可以取得很好的性能。He 等人[13]利用空间金字塔池化方法，在最后一个卷积层中加入一个池化层来对输出的特征进行池化，实现了卷积神经网络的输入大小非固定尺度。Wang 等人[14]通过构造一个 3 维卷积核最大池化的网络，实现对 RGB-D 视频的自动识别。Wang 等人[15]在对主流的一些网络结构进行调整，提出非常深的双流卷积神经网络并应用于视频的动作识别中。王忠民等人[16]利用卷积神经网络结合 SVM 支持向量机对智能终端采集的五种日常人体动作进行识别。韩敏捷[17]提出 2 模态动作识别方法，对于 Kinect 传感器捕获的静态信息使用卷积神经网络处理，动态信息则用递归神经网络，最后融合两种模型提取的特征进行动作识别。

浅层学习网络，在训练样本比较有限的条件下，表示复杂函数的能力有限，且模型的泛化能力也有很大的局限性。Simonyan 等人[18]在大规模数据集中验证了当卷积神经网络的深度增加到 16 至 19 个权重层时，识别的结果有很大程度的改善。GoogLeNet 网络[19]是在传统深度卷积神经网络[20]的基础上加入多个 inception 网络模型的结构。本文提出批归一化变换与 GoogLeNet 网络模型相结合的网络架构并应用到视频人体动作识别领域，相对于传统的深度卷积神经网络在训练算法及网络结构两方面进行改进。空间流网络通过视频帧的 RGB 图像来获取运动的外观信息，而时间流则是通过连续帧间的光流场来捕获运动信息，最后将时空网络融合既考虑外观信息又关注到运动信息，实现提高动作识别准确率的目的。本文还探究了 Dropout 层不同的 dropout 率以及时空网络不同线性加权融合比例对动作识别准确率的影响。

1 网络架构改进

深度神经网络在训练时，各层网络的输入分布会受到上一层参数的影响，随着网络层数的叠加，网络层的微小变动所产生的影响就会不断被放大，这就有可能会产生梯度消失或者梯度爆炸问题。随着网络层的参数被不断更新，各层的输入范围也会有所差异和变化，这会导致网络的收敛速度减慢，整个网络有可能会收敛于一个不理想的局部最优值。以上问题的出现都是由于内部协变量迁移(internal covariate shift)[21]引起的。而要消除内部协变量迁移所带来的副作用，可以通过修改网络结构，或者在激活层中加入白化(whitening)处理,也可以改变参数调优算法[22～25]。为解决上述问题，本文借鉴文献[26]在 ImageNet 图像分类领域上提出的批归一化(batch normalization)算法处理一些网络层输入的思想应用到动作识别领域中，对视频动作训练样本的网络输入进行微批量(mini-batch)归一化处理。

1.1 批归一化处理算法

传统的批归一化如式(1)所示。

$$\hat{X} = norm(x, X) \tag{1}$$

其中：x 表示网络中某一层的输入矢量；$X = \{x_{1,\dots,N}\}$ 表示整个训练集的输入集合。从式(1)可以看出，批归一化的输出取决于输入 x 和整个训练样本的取值 X。在训练集 X 中，每层网络的输入 x 是由上一层网络的输出生成，x 受模型参数的影响。因此在用反向传播算法更新网络参数的过程中，需要计算与批归一化相应的 x 和 X 的雅克比矩阵，如式(2)所示。

$$\frac{\partial norm(x, X)}{\partial x} \quad \frac{\partial norm(x, X)}{\partial X} \tag{2}$$

如果对每层的输入都加入批归一化处理，会非常耗时(需要计算协方差矩阵)。对此 Ioffe 等人[26]对传统的批归一化算法提出两点简化改进：

a)简化改进，是对输入的各维进行独立的批归一化处理而不是联合归一化处理,如式(3)所示.

$$\hat{X}^{(k)} = \frac{x_i^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{var}[x^{(k)}]}} \tag{3}$$

其中: $x^{(k)}$ 表示输入样本的第 k 维; $\mathrm{E}[x^{(k)}]$ 、 $\mathrm{var}[x^{(k)}]$ 分别表示输入的期望和方差。Lécun 等人[27]证明了，即使训练特征是不相关的，式(3)批归一化算法也可以加速收敛，但是它可能会改变各层原来的表示，使得输入无法完整表达原有的输出特征。因此，为了保证引入的批归一化变换是恒等式，需要对每个输入 $x^{(k)}$ 加入一对参数 $\lambda^{(k)}, \beta^{(k)}$ ，如式(4)所示.

$$y^{(k)} = \lambda^{(k)} \hat{X}^{(k)} + \beta^{(k)} \tag{4}$$

其中：$\lambda^{(k)} = \mathrm{var}[x^{(k)}]$ 表示输入的标准差，相当于对输入 $x^{(k)}$ 进行尺度变换；$\beta^{(k)} = \mathrm{E}[x^{(k)}]$ 相当于对 $x^{(k)}$ 进行平移变换。这两个参数和神经网络模型中的参数一样通过训练学习获得，用来恢复模型的表达能力。

b)简化改进，是在随机梯度训练中采用微批量(mini-batch)样本进行训练，在每个微批量样本上对每层进行计算，估计该层的均值和方差，因此在批归一化处理中计算的神经网络统计量(方差和均值)可以用于梯度反向传播中。假定微批量样本 B 的大小是 m ，某层的输入某维是 x ，则逐维归一化如式(5)所示。

$$BN_{\lambda,\beta} : x_{1,\dots,m} \rightarrow y_{1,\dots,m} \tag{5}$$

下面在算法 1 中介绍了在深度卷积神经网络的某一层中插入归一化变换算法，相对于未加入批归一化处理的网络输入是 x ，加入之后的输入变为 $BN(x)$ 。算法 2 是对整个深度卷积神经网络插入批归一化变换的算法流程。由算法 1 可以看出，归一化处理包括对输入进行归一化以及对于归一化后的数据进行尺度不变的平移变换。

算法 1　微批量归一化变换算法

输入：微批量输入值：$B = \{x_{1,\dots,m}\}$ ；

待训练学习的参数：λ, β 。

输出：$y_i = BN_{\lambda,\beta}(x_i)$ 。

开始：

$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ 　//计算微批量样本 B 的均值

$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$ 　//计算 B 的方差

$\hat{X}_i \leftarrow \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}}$ //归一化处理，ε 是一个常量

$y_i \leftarrow \lambda\hat{X}_i+\beta \equiv BN_{\lambda,\beta}(x_i)$ //尺度不变平移变换

结束

算法 2 神经网络的批归一化变换算法

输入：神经网络 N，训练参数集合 θ；

每层网络输入集合 $\{x^k\}_{k=1}^K$。

输出：批归一化处理后的网络 $N_{BN}^{\inf}$。

开始：

$N_{BN}^{\inf} \leftarrow N$

a) $for \quad k=1,\ldots,K \quad do$ //开始训练

b) 在 $N_{BN}^{\inf}$ 的基础上进行仿射变换：

$y^{(k)}=BN_{\lambda^{(k)},\beta^{(k)}}(x^{(k)})$

c) 对 $N_{BN}^{\inf}$ 网络的每一层：

用仿射变换的输入 $y^{(k)}$ 替代原输入 $x^{(k)}$

d) $end \quad for$

e) 训练 $N_{BN}^{\inf}$，更新网络参数：$\theta \cup \{\lambda^{(k)},\beta^{(k)}\}_{k=1}^K$

f) 冻结参数，推出批归一化后的网络 $N_{BN}^{\inf}$：$N_{BN}^{\inf} \leftarrow N_{BN}^{tr}$

g) $for \quad k=1,\ldots K \quad do$

//令 $x \equiv x^{(k)}; \lambda \equiv \lambda^{(k)}; \mu_B \equiv \mu_B^{(k)}$

h) 对每个微批量样本 B（大小为 m）进行训练，然后计算 B 的均值和方差：

$$\mathrm{E}[x] \leftarrow \mathrm{E}_B[\mu_B]; \mathrm{var}[x] \leftarrow \frac{m}{m-1}\mathrm{E}_B[\sigma_B^2]$$

i) 在 $N_{BN}^{\inf}$ 中，用下面公式替代原有的 BN 变换 $y=BN_{\lambda,\beta}(x)$：

$$y=\frac{\lambda}{\sqrt{\mathrm{var}[x]+\varepsilon}}\cdot(x)+(\beta-\frac{\lambda\mathrm{E}[x]}{\sqrt{\mathrm{var}[x]+\varepsilon}})$$

j) $end \quad for$

结束

加入归一化处理后的深度卷积神经网络在训练中需要使用反向传播算法来计算损失函数 l 的梯度，同时还需要计算批归一化变换中加入的参数。式(6)给出了用反向传播算法求解网络参数梯度的过程。

$$\begin{aligned}
&\frac{\partial l}{\partial \hat{X}_i}=\frac{\partial l}{\partial y_i}\lambda \\
&\frac{\partial l}{\partial \sigma_B^2}=\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{X}_i}\cdot(x_i-\mu_B)\cdot(-\frac{1}{2})(\sigma_B^2+\varepsilon)^{(-\frac{3}{2})} \\
&\frac{\partial l}{\partial \mu_B}=(\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{X}_i}\cdot\frac{-1}{\sqrt{\sigma_B^2+\varepsilon}})+\frac{\partial l}{\partial \sigma_B^2}\cdot\frac{\sum_{i=1}^{m}-2(x_i-\mu_B)}{m} \\
&\frac{\partial l}{\partial x_i}=\frac{\partial l}{\partial \hat{X}_i}\cdot\frac{1}{\sqrt{\sigma_B^2+\varepsilon}}+\frac{\partial l}{\partial \sigma_B^2}\cdot\frac{2(x_i-\mu_B)}{m}+\frac{\partial l}{\partial \mu_B}\cdot\frac{1}{m} \\
&\frac{\partial l}{\partial \lambda}=\sum_{i=1}^{m}\frac{\partial l}{\partial y_i}\hat{X}_i \\
&\frac{\partial l}{\partial \beta}=\sum_{i=1}^{m}\frac{\partial l}{\partial y_i}
\end{aligned} \tag{6}$$

1.2 批归一化与 GoogLeNet 相结合的网络构建

本文提出批归一化变换与 GoogLeNet 网络模型相结合的网络结构，运用到视频人体动作识别中，具体的处理过程是对每个卷积层的输入特征进行批归一化处理，然后将批归一化处理后的特征输入到激活函数 ReLU 层中。图 1 所示为 GoogLeNet 中的一个 inception 层的批归一化处理后的 inception 网络结构。

整个改进的网络模型除了如图 1 所示的在 inception 网络模型中的每个卷积层后面都加入批归一化处理层之外，在底层网络中的每一个卷积层的后面都跟随有一个批归一化处理层，在批归一化处理后同样是接入到 ReLU 激活层，再接入后续的网络中。本文应用到人体动作识别的深度卷积神经网络的结构如表 1 所示。

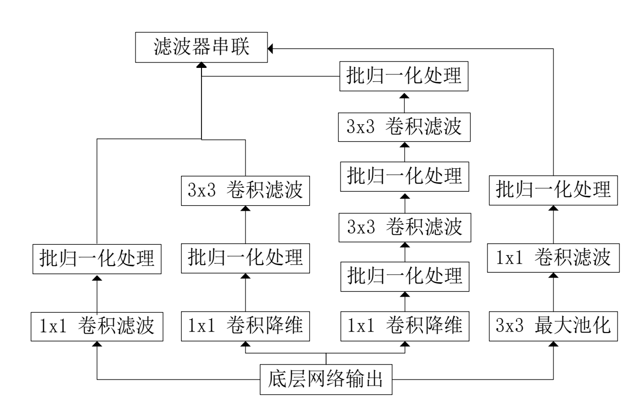


图 1 加入批归一化处理的 inception 网络结构

2 时空双流网络构建

2.1 双流网络

视频可以看做是由时间和空间两部分组成。在空间部分，每个独立的帧都包含有场景和物体的外观信息；在时间部分，则包括相机和物体的运动信息。时空双流网络模型如图 2 所示。空间流网络的输入是 RGB 图像，时间流的输入是光流场。

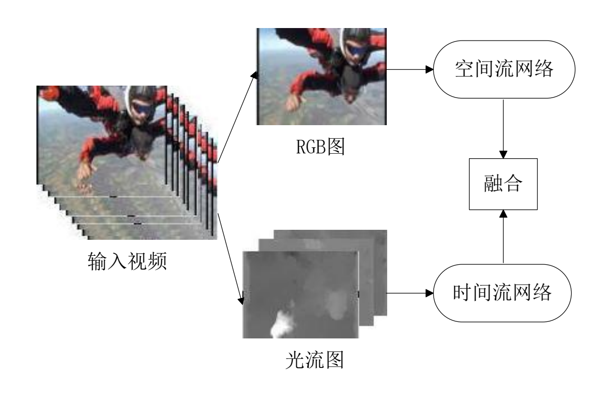


图 2 深度时空双流网络模型结构

2.2 网络训练

公开的动作识别数据集相对于 ImageNet[28]数据集而言，数据量比较小。当卷积神经网络比较深时，训练集较小容易使网络陷入过拟合现象。因此，先进行一些预处理，利用数据增强技术来扩充训练集，通过对网络进行预训练处理来进行网络权重的初始化。

数据增强是对训练数据集进行几何变换达到增加训练集的过程。由于随机裁剪技术比较倾向于选中图像的中心区块，易造成过拟合，所以对图像帧的边角和中心区块进行裁剪，增强

尺度多样性。具体的做法是将输入的数据大小固定为256x340，然后随机从集合{256,224,192}中选择一个候选的裁剪大小进行裁剪，最后在把裁剪下来的区块调整为224x224大小。

预训练。将所用的深度卷积神经网络在ImageNet数据集上进行预训练处理。由于空间网络的输入是RGB图像，所以可直接在ImageNet上进行网络预训练，而时间网络的输入是10帧堆叠的光流场，需要进行一些网络的调整。本文使用的是TV-L1光流[29]，首先利用opencv提取动作视频的光流场；然后通过线性变换将光流离散到[0,255]区间,保证与RGB同区间；最后将在ImageNet预训练的空间网络模型的第一层的滤波器在通道中做平均，将取平均后的结果进行复制20次(垂直和水平方向的光流)，作为时间网络的初始化。

表1　动作识别网络结构

类型	滑窗大小输出/步长	输出维度	Inception 第1卷积支路滑窗大小(卷积输出)	Inception 第2卷积支路滑窗大小(卷积输出)		Inception 第3卷积支路滑窗大小(卷积输出)			Inception 第4池化支路滑窗大小（池化输出|卷积输出）	
卷积	7x7/2	112x112x64								
最大池化	3x3/2	56x56x64								
卷积	3x3/1	56x56x192								
最大池化	3x3/2	28x28x192								
Inception 1		28x28x256	1x1 64	1x1 64	3x3 64	1x1 64	3x3 96	3x3 96	3x3	1x1 32
Inception 2		28x28x320	1x1 64	1x1 64	3x3 96	1x1 64	3x3 96	3x3 96	3x3	1x1 64
Inception 3		14x14x576		1x1 128	3x3 160	1x1 64	3x3 96	3x3 96	3x3	
Inception 4		14x14x576	1x1 224	1x1 64	3x3 96	1x1 96	3x3 128	3x3 128	3x3	1x1 128
Inception 5		14x14x576	1x1 192	1x1 96	3x3 128	1x1 96	3x3 128	3x3 128	3x3	1x1 128
Inception 6		14x14x608	1x1 160	1x1 128	3x3 160	1x1 128	3x3 160	3x3 160	3x3	1x1 128
Inception 7		14x14x608	1x1 96	1x1 128	3x3 192	1x1 160	1x1 192	3x3 192	3x3	1x1 128
Inception 8		14x14x1056		1x1 128	3x3 192	1x1 192	3x3 256	3x3 256	3x3	
Inception 9		7x7x1024	1x1 352	1x1 192	3x3 320	1x1 160	3x3 224	3x3 224	3x3	1x1 128
Inception 10		7x7x1024	1x1 352	1x1 192	3x3 320	1x1 192	3x3 224	3x3 224	3x3	1x1 128
平均池化	7x7x1	1x1x1024								
dropout		1x1x1024								
全连接		1x1x101								
Softmax		1x1x101								

网络训练。因为在ImageNet上做了预训练，所以在训练时要使用更小的学习率。动量值设为0.9。对于空间网络，基础学习率为0.001，每迭代1 800次，则降为原来的1/10，最大迭代次数为5 000。时间网络的学习率为0.003，迭代至15 000次，学习率降为原来的1/10，迭代至18 000次，学习率再降1/10，迭代至20 000次，网络训练结束。

3　实验

3.1　实验数据

本文的实验数据是采用公开的视频动作识别数据集UCF101[30]和HMDB51[31]。部分动作的示意图如图3所示。

UCF101数据集包含有101类动作，共有13 320个视频段。UCF101数据集是由在无约束的现实环境下拍摄的网络视频构成，视频帧像素比较低，包含有不同的光照信息，存在部分遮挡和相机运动的情况。该数据集将动作划分为五种类型：a)人与人交互类，如剪头发、头部按摩等5个类别；b)演奏乐器类，如吹笛子、拉小提琴等10个类别；c)仅含人体运动类，如吹蜡烛、打太极等16类；d)人与物交互类，如吹头发、切菜等20个类别；e)运动类，如打台球、蛙泳等50个类别。

HMDB51数据集包含有51类动作，共有6 849个视频段。HMDB51数据集大部分来源于电影片段，小部分来自YouTube等视频网站。同样，HMDB51也被划分为五种类型：a)与物体

交互的面部表情类，如抽烟、喝水等 3 个类别；b)一般的面部动作类，如微笑、说话动作等 4 个类别；c)人与人交互的身体运动类，如拥抱、亲吻等 7 个类别；d)人与物交互的身体运动类，如拔剑、骑马等 18 个类别；e)一般的身体运动类，如鼓掌、倒立等 19 个类别。



图 3　部分动作示意图

3.2　实验结果与分析

在 Linux 系统下搭建的 caffe 平台单 GPU 上进行实验。按文献[30,31]规定的数据集分割标准，使用三种训练/分类分割(train/test split)方法。其中每种分割(split)方式，都是将数据集大约分为 70%训练集和 30%测试集。

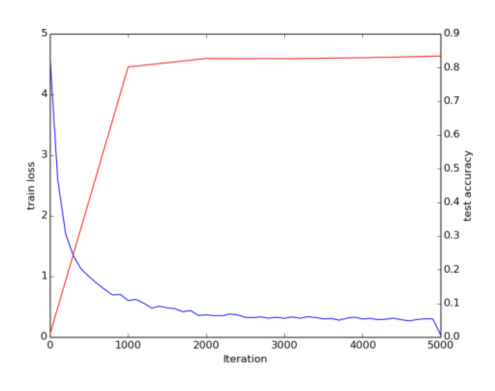
一般情况下，可以在深度卷积神经网络中加入 Dropout 层来避免过拟合[32]。本文探究了在新构建的时空动作识别网络中 Dropout 层的 dropout 率(dropout_ratio)参数值对识别准确率的影响。用不同的 dropout 率参数值在 UCF101 数据集的分割方式 1(split1)下进行实验分析，结果如表 2 所示。

表 2　不同 dropout 率对识别准确率的影响

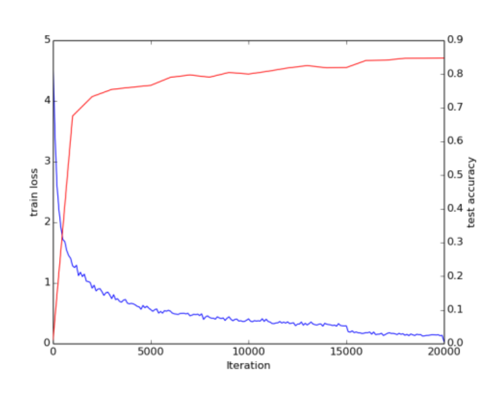
网络	(dropout 率)准确率	(dropout 率)准确率	(dropout 率)准确率
时间网络	(0.4)86.56%	(0.6)86.48%	(0.7)86.78%
空间网络	(0.4)82.61%	(0.6)83.16%	(0.8)83.68%

表 2 展示了 dropout_ratio 参数的不同数值在 UCF101 的 split1 数据集上的动作识别准确率。从表 2 可以看出，时间网络的 dropout 率为 0.7 时，比 0.4 和 0.6 的识别率要分别高出 0.22%和 0.3%；空间网络的 dropout 率为 0.8 时，比 0.4 和 0.6 的识别率要分别高出 1.07%和 0.52%。本文在后续的实验中分别将时间网络和空间网络的 dropout 率设置为 0.7 和 0.8。

图 4 展示了 UCF101 数据集在分割方式三(split3)下的时空网络训练迭代收敛图。从图 4 (a)中可以看出，在空间流上，当训练迭代次数达到 1 000 时，accuracy 值接近 80%，train 的 loss 值迅速减小，之后 accuracy 曲线慢慢上升，loss 曲线慢慢下降；当训练迭代到 2 000 次以后，accuracy 保持在 80%以上，loss 值保持在 0.5 以下，随着迭代的进行，收敛情况趋于稳定。从图 4(b)中可以看出，在时间流上，当训练迭代次数达到 1 500 时，train 的 loss 值迅速减小，之后 accuracy 曲线慢慢上升，loss 曲线慢慢下降；当训练迭代到 15 000 次以后，accuracy 保持在 80%以上，train 的 loss 值保持在 0.3 以下，随着迭代的进行，收敛情况趋于稳定。



(a)空间网络训练迭代收敛图



(b)时间流网络训练迭代收敛图

图 4　时空网络的训练迭代收敛图

改进的时空网络架构在 UCF101 和 HMDB51 数据集上的动作识别准确率记录在表 3 中。将时空网络分类结果用线性加权的方式进行融合[15]。本文还探究了网络识别置信度的不同比值对动作识别准确率的影响，得到时空融合网络的识别率如表 4 所示。最后将本文方法与现有的一些实验方法进行比较和分析，比较的结果如表 5 所示。

表 3　改进的时空网络识别准确率

网络	分割方式	UCF101/%	HMDB51/%
空间网络	split1	83.68	53.99
	split2	81.76	48.69
	split3	83.75	49.67
	取平均	83.06	50.78
时间网络	split1	86.78	62.81
	split2	89.91	61.90
	split3	89.73	65.42
	取平均	88.81	63.38

表 3 展示了空间流和时间流深度卷积神经网络在 UCF101

和 HMDB51 数据集上对于三种不同分割方式下的识别准确率。从表中可以看出，时间流网络提取的运动信息比空间流网络上提取的外观信息具有更高的识别率，这也说明了对于动作识别任务，运动信息比外观信息更为重要。

表 4　时空融合网络识别准确率

分割方式	UCF101/%			HMDB51/%		
空间：时间	1:1	1:1.2	1:1.5	1:1	1:1.2	1:1.5
split1	93.37	92.31	92.12	69.48	69.22	69.08
split2	93.95	93.97	93.82	67.32	67.35	67.06
split3	93.17	93.17	93.29	68.17	67.91	67.71
取平均	93.50	93.15	93.08	68.32	68.16	67.95

表 4 展示了融合后的时空网络的识别准确率。对于每一种分割方式下实验得到的时间网络上的识别率和空间网络上的识别率，进行线性加权融合得到最终的识别率。由表 4 可以看出，空间网络和时间网络分类的识别置信度的权值设置为 1:1 时，融合的双流卷积神经网络的识别性能要优于 1:1.2 和 1:1.5 的情况。对比表 3 可以看出，在动作识别任务中，融合的时空双流深度卷积神经网络能有效改善单独的网络在识别上的准确率。

表 5　不同算法识别准确率的比较

方法	UCF101	HMDB51
Improved dense trajectories[4][33]	85.9%	57.2%
IDT with higher-dimensional encoding[34]	87.9%	61.1%
Two-stream[15]	88%	59.4%
Very deep two-stream[18]	91.4%	
KVMF[35]	93.1%	63.3%
本文方法	93.50%	68.32%

表 5 给出了本文方法和动作识别中比较典型的动作识别方法在 UCF101 和 HMDB51 数据集上的识别准确率的对比。Improved dense trajectories[4, 33]都是使用密集轨迹算法，IDT with higher-dimensional encoding[34]是对 BOVW 视觉词袋模型进行改进融合多维特征实现更高维特征编码，这两种动作识别的方法都是比较传统的手工设计特征的方法。Two-stream[15]是通过构建一个双流时空网络模型来对动作识别，但是网络比较浅层。Very deep two-stream[18]非常深的网络架构，在深度上对网络进行改进。KVMF 算法[35]通过对视频段截取多个 3D volumes 来作为网络的输入，用每个 volume 得到的预测向量来表示所属动作的类别概率。由表 5 可以看出，本文提出的改进的融合时空双流深度卷积神经网络在该数据集上具有更好的动作识别能力。

4　结束语

本文在人体动作识别任务上，提出改进的深度卷积神经网络模型结构，并利用改进后的网络模型构建时空双流深度卷积神经网络架构。在 ImageNet 数据集上进行微调，融合的深度卷积神经网络在 UCF101 和 HMDB51 数据集上分别取得了 93.50% 和 68.32%的识别率。目前深度卷积神经网络算法已经成功应用在模式识别等领域的实验研究中，但是与实时响应的商业化应用还有一段距离，主要是因为训练网络需要耗费很长的时间，所以今后可以在并行计算深度卷积神经网络算法方面做深入研究。

参考文献：

[1] Jhuang H, Serre T, Wolf L, et al. A biologically inspired system for action recognition [C]// Proc of IEEE International Conference on Computer Vision. 2007: 1-8.

[2] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C]// Proc of Computer Vision and Pattern Recognition. 2014: 1725-1732.

[3] 凌佩佩，邱崧，蔡茗名，等. 结合特权信息的人体动作识别 [J]. 中国图象图形学报, 2017, 22 (4): 482-491.

[4] Wang H, Schmid C. Action recognition with improved trajectories [C]// Proc of IEEE International Conference on Computer Vision. 2013: 3551-3558.

[5] 申晓霞，张桦，高赞，等. 基于深度信息和 RGB 图像的行为识别算法 [J]. 模式识别与人工智能, 2013, 26 (8): 722-218.

[6] 张杰，吴剑章，汤嘉立，等. 基于时空图像分割和交互区域检测的人体动作识别方法 [J]. 计算机应用研究, 2017, 34 (1): 302-305, 320.

[7] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images [C]// Proc of Computer Vision and Pattern Recognition. 2011: 1297-1304.

[8] Ofli F, Chaudhry R, Kurillo G, et a1. Sequence of the most informative j oints (SMIJ): a new representation for human skeletal action recognition [J]. Journal of Visual Communication and Image Representation, 2014, 25 (1): 24-38.

[9] Chen C, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns [C]// Proc of Applications of Computer Vision. 2015: 1092-1099.

[10] 赵晓健，曾晓勤. 基于稠密光流轨迹和稀疏编码算法的行为识别方法 [J]. 计算机应用, 2016, 36 (1): 181-187.

[11] 李亚玮，金立左，孙长银，等. 基于光流约束自编码器的动作识别 [J]. 东南大学学报：自然科学版, 2017, 47 (4): 691-696.

[12] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1 (4): 568-576.

[13] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 37 (9): 1904-1916.

[14] Wang K, Wang X, Lin L, et al. 3D human activity recognition with reconfigurable convolutional neural networks [C]// Proc of ACM International Conference on Multimedia. 2015: 97-106.

[15] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-

stream ConvNets [J]. arXiv preprint arXiv: 1507. 02159, 2015.

[16] 王忠民，曹洪江，范琳. 一种基于卷积神经网络深度学习的人体行为识别方法 [J]. 计算机科学, 2016, 43 (s2): 56-58.

[17] 韩敏捷. 基于深度学习框架的多模态动作识别 [J]. 计算机与现代化, 2017 (7): 48-52.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409. 1556, 2014.

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C]// Proc of Computer Vision and Pattern Recognition. 2015.

[20] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1 (4): 541-551.

[21] Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function [J]. Journal of Statistical Planning and Inference, 2000, 90 (2): 227-244.

[22] Wiesler S, Richard A, Schluter R, et al. Mean-normalized stochastic gradient for large-scale deep learning [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2014: 180-184.

[23] Raiko T, Valpola H, Lecun Y. Deep learning made easier by linear transformations in perceptrons [C]// Proc of the 15th Internation Conference on Artificial Intelligence and Statistics. 2012, 22: 924-932.

[24] Povey D, Zhang X, Khudanpur S. Parallel training of deep neural networks with natural gradient and parameter averaging [J]. arXiv preprint arXiv: 1410. 7455, 2014.

[25] Desjardins G, Simonyan K, Pascanu R, et al. Natural neural networks [J]. Computer Science, 2015, 22 (8): 847-856.

[26] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// Proc of the 32nd International Conference on Machine Learning. 2015: 448-456.

[27] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[28] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C]// Proc of Computer Vision and Pattern Recognition. 2009: 248-255.

[29] Pérez J S. TV-L1 optical flow estimation [J]. Image Processing on Line, 2013, 2 (4): 137-150.

[30] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [J]. arXiv preprint arXiv: 1212. 0402, 2012.

[31] Kuehne H, Jhuang H, Stiefelhagen R, et al. HMDB51: a large video database for human motion recognition [C]// Proc of IEEE International Conference on Computer Vision. 2012: 2556-2563.

[32] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3 (4): 212-223.

[33] Wang H, Schmid C. LEAR-INRIA submission for the thumos workshop [C]// Proc of ICCV Workshop on THUMOS Challenge. 2013.

[34] Peng X, Wang L, Wang X, et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice [J]. Computer Vision and Image Understanding, 2016, 150 (C): 109-125.

[35] Zhu W, Hu J, Sun G, et al. A key volume mining deep framework for action recognition [C]// Proc of Computer Vision and Pattern Recognition. 2016: 1991-1999.